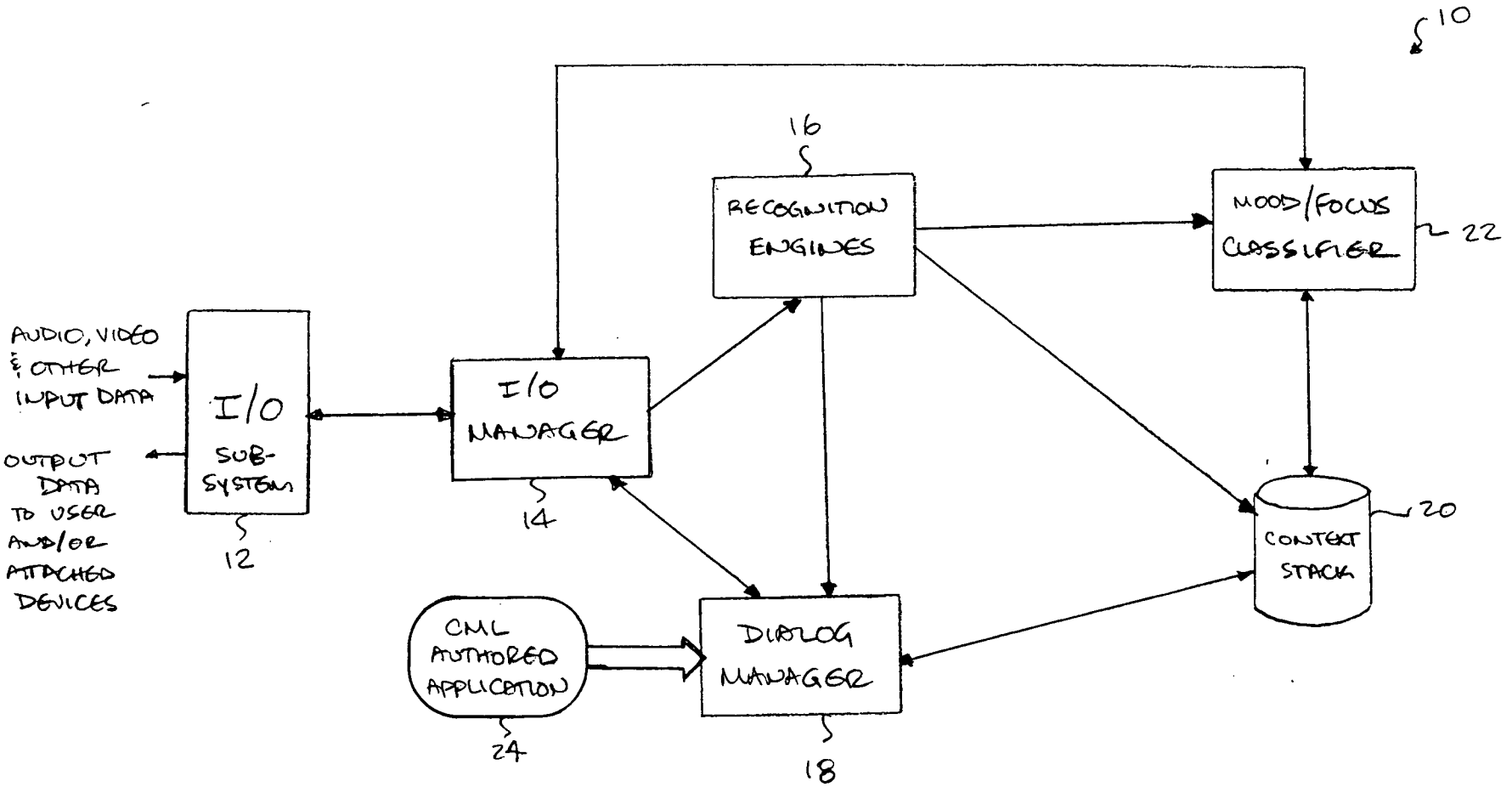
10
16
RECOGNITION ENGINES
MOOD/FOCUS CLASSIFIER
22
AUDIO, VIDEO & OTHER INPUT DATA
OUTPUT DATA TO USER AND/OR ATTACHED DEVICES
I/O SUB-SYSTEM
12
I/O MANAGER
14
CONTEXT STACK
20
CML AUTHORED APPLICATION
24
DIALOG MANAGER
18

FIG. 1

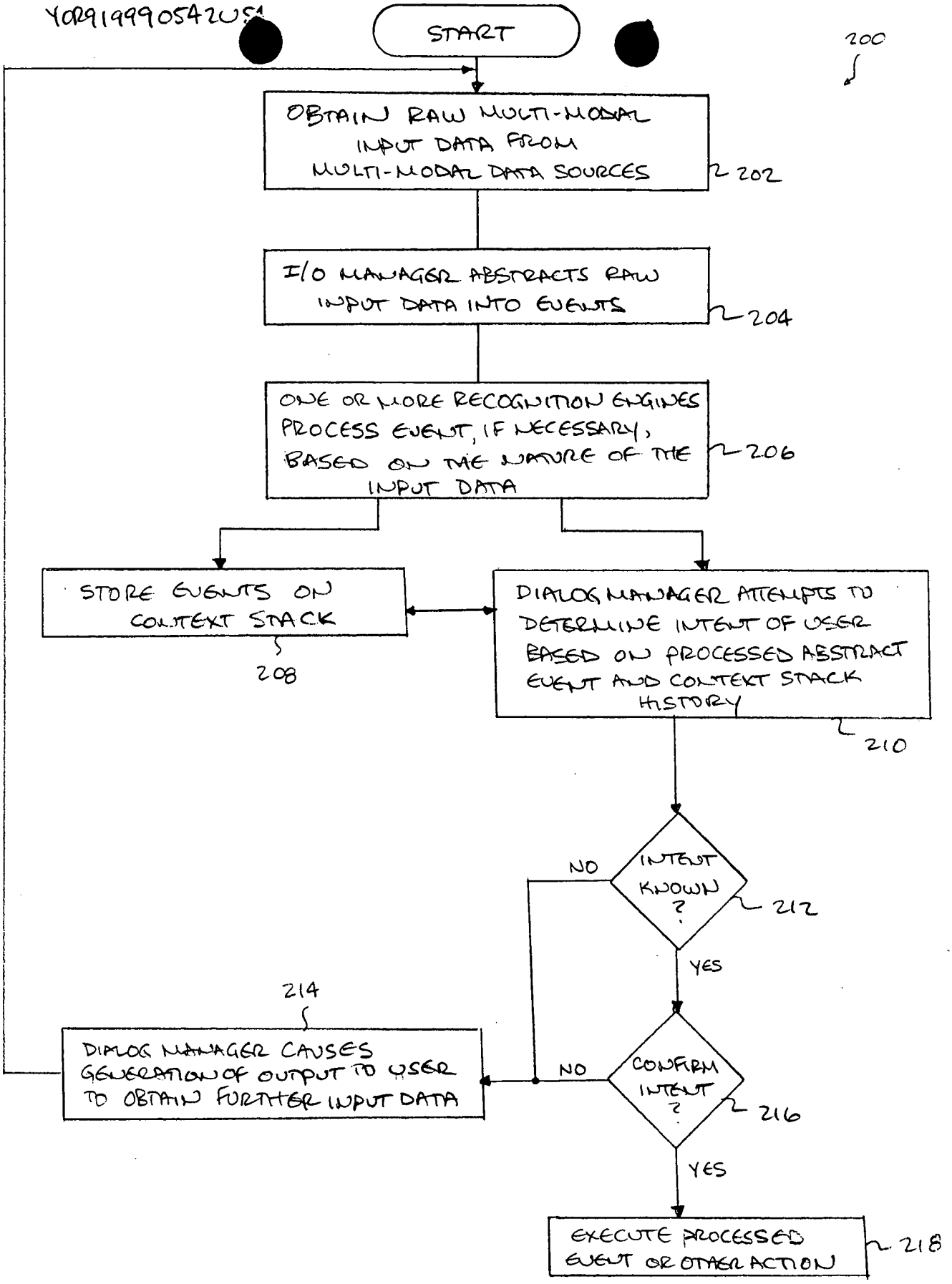
START
200
OBTAIN RAW MULTI-MODAL INPUT DATA FROM MULTI-MODAL DATA SOURCES
202
I/O MANAGER ABSTRACTS RAW INPUT DATA INTO EVENTS
204
ONE OR MORE RECOGNITION ENGINES PROCESS EVENT, IF NECESSARY, BASED ON THE NATURE OF THE INPUT DATA
206
STORE EVENTS ON CONTEXT STACK
208
DIALOG MANAGER ATTEMPTS TO DETERMINE INTENT OF USER BASED ON PROCESSED ABSTRACT EVENT AND CONTEXT STACK HISTORY
210
INTENT KNOWN ?
212
NO
YES
214
DIALOG MANAGER CAUSES GENERATION OF OUTPUT TO USER TO OBTAIN FURTHER INPUT DATA
CONFIRM INTENT ?
216
NO
YES
EXECUTE PROCESSED EVENT OR OTHER ACTION
218

FIG. 2

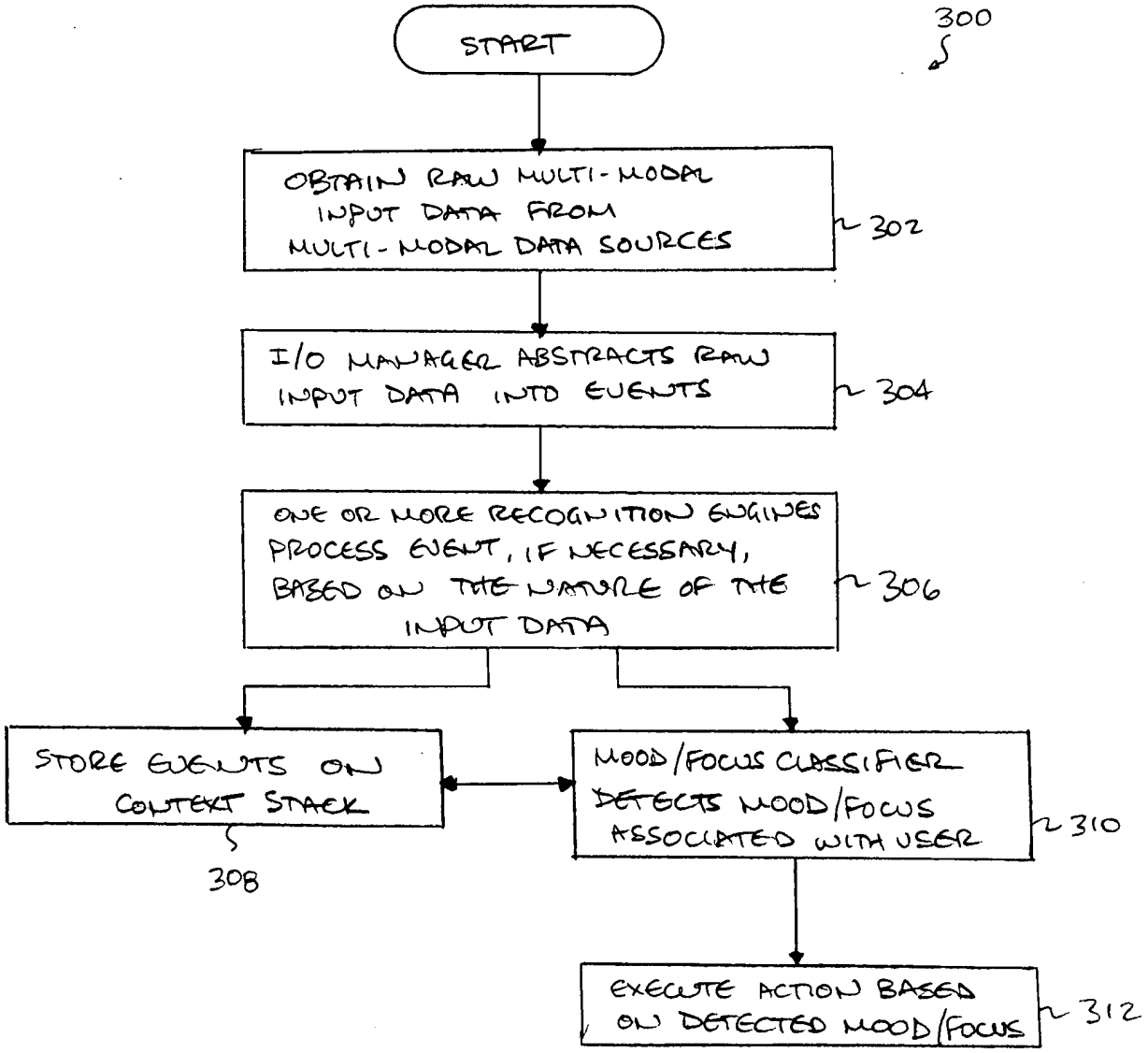
START
300
OBTAIN RAW MULTI-MODAL INPUT DATA FROM MULTI-MODAL DATA SOURCES
302
I/O MANAGER ABSTRACTS RAW INPUT DATA INTO EVENTS
304
ONE OR MORE RECOGNITION ENGINES PROCESS EVENT, IF NECESSARY, BASED ON THE NATURE OF THE INPUT DATA
306
STORE EVENTS ON CONTEXT STACK
308
MOOD/FOCUS CLASSIFIER DETECTS MOOD/FOCUS ASSOCIATED WITH USER
310
EXECUTE ACTION BASED ON DETECTED MOOD/FOCUS
312

FIG. 3

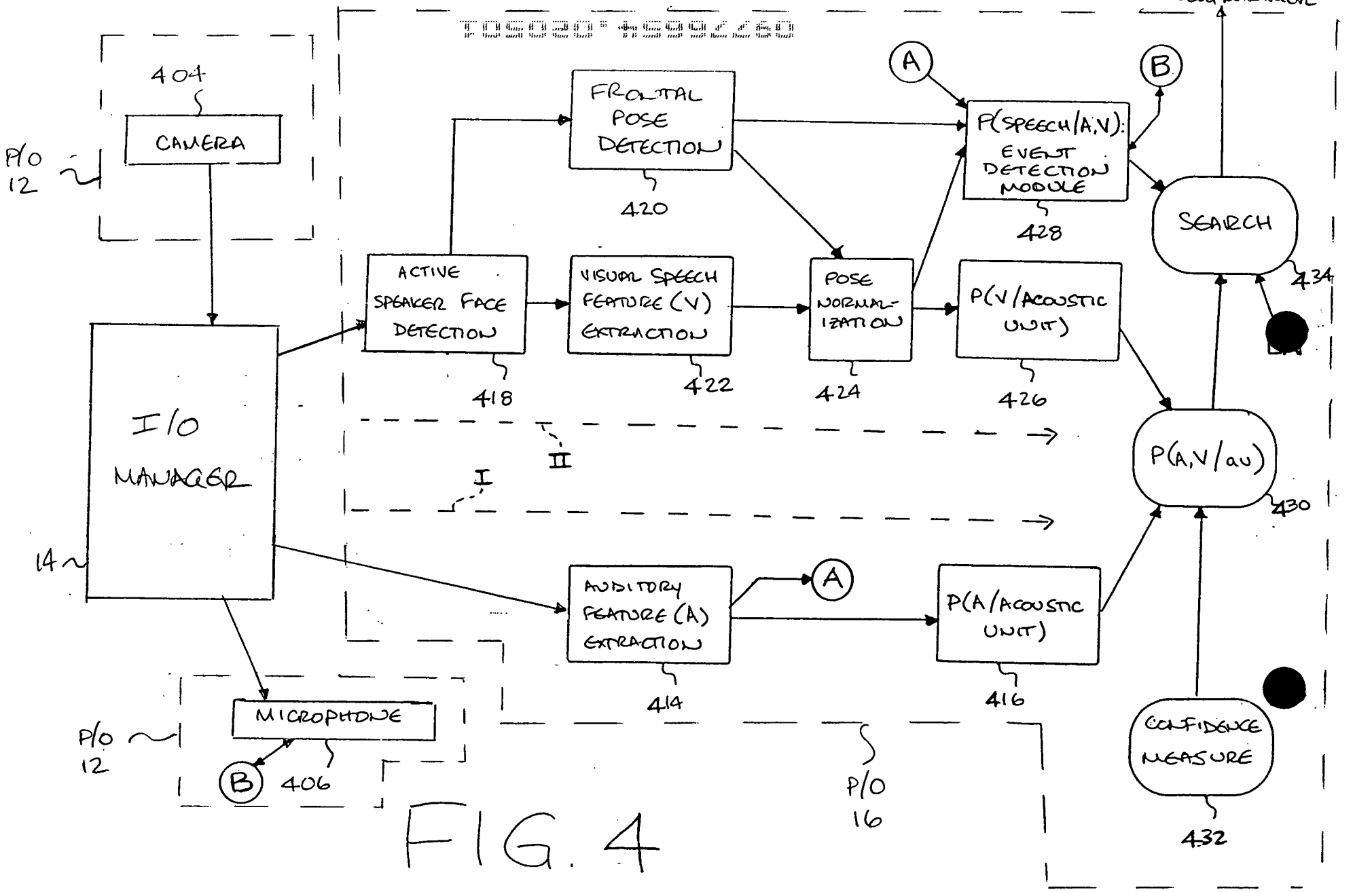
TO DIALOG MANAGER
404
CAMERA
P/O 12
I/O MANAGER
14
ACTIVE SPEAKER FACE DETECTION
418
FRONTAL POSE DETECTION
420
VISUAL SPEECH FEATURE (V) EXTRACTION
422
POSE NORMALIZATION
424
P(V/ACOUSTIC UNIT)
426
P(SPEECH/A,V): EVENT DETECTION MODULE
428
SEARCH
434
P(A,V/au)
430
CONFIDENCE MEASURE
432
AUDITORY FEATURE (A) EXTRACTION
414
P(A/ACOUSTIC UNIT)
416
MICROPHONE
406
P/O 12
P/O 16
A
B
I
II

FIG. 4

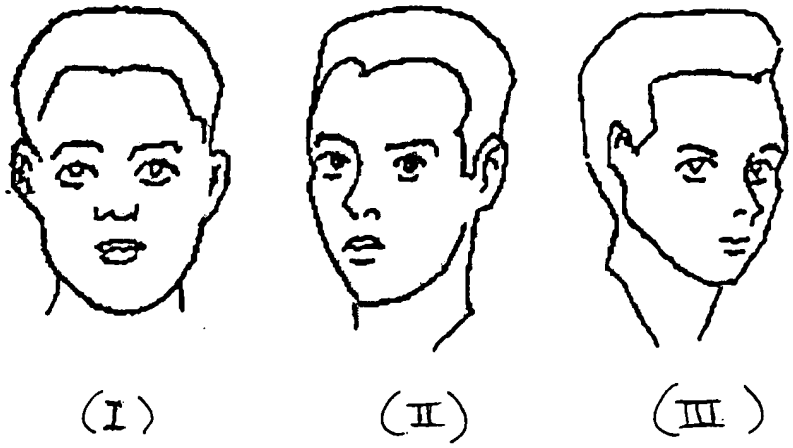

FIG. 5A

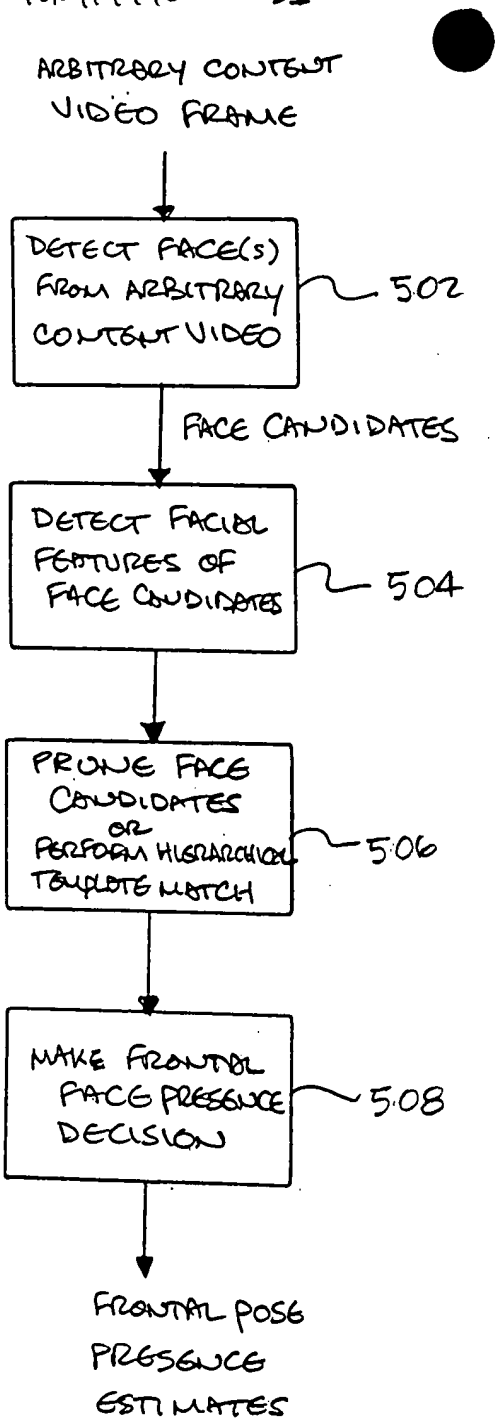
ARBITRARY CONTENT VIDEO FRAME
DETECT FACE(S) FROM ARBITRARY CONTENT VIDEO
502
FACE CANDIDATES
DETECT FACIAL FEATURES OF FACE CANDIDATES
504
PRUNE FACE CANDIDATES OR PERFORM HIERARCHICAL TEMPLATE MATCH
506
MAKE FRONTAL FACE PRESENCE DECISION
508
FRONTAL POSE PRESENCE ESTIMATES

FIG. 5B

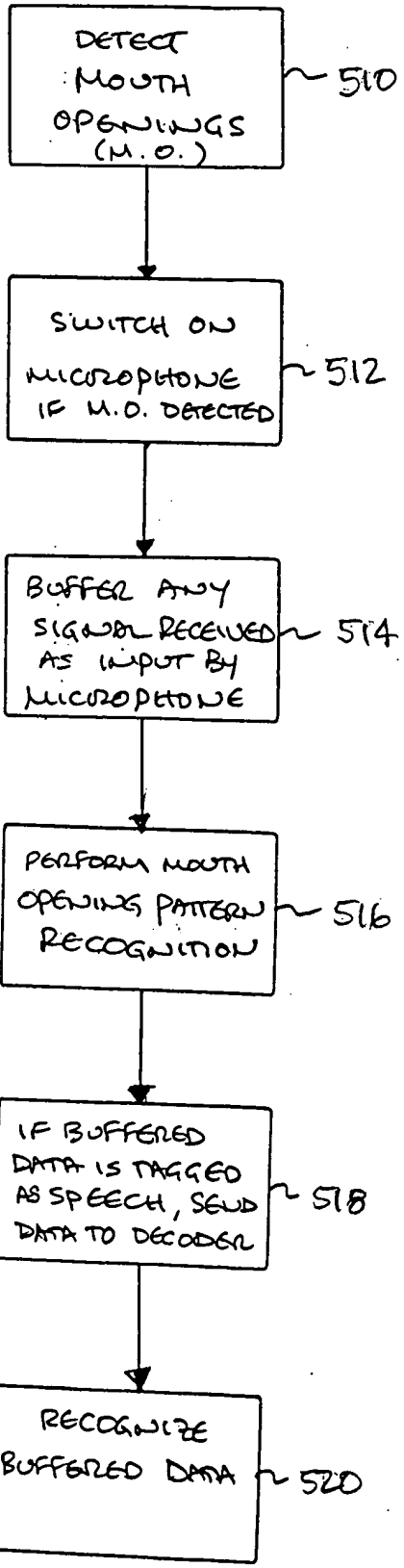
DETECT MOUTH OPENINGS (M.O.)
510
SWITCH ON MICROPHONE IF M.O. DETECTED
512
BUFFER ANY SIGNAL RECEIVED AS INPUT BY MICROPHONE
514
PERFORM MOUTH OPENING PATTERN RECOGNITION
516
IF BUFFERED DATA IS TAGGED AS SPEECH, SEND DATA TO DECODER
518
RECOGNIZE BUFFERED DATA
520

FIG. 5C

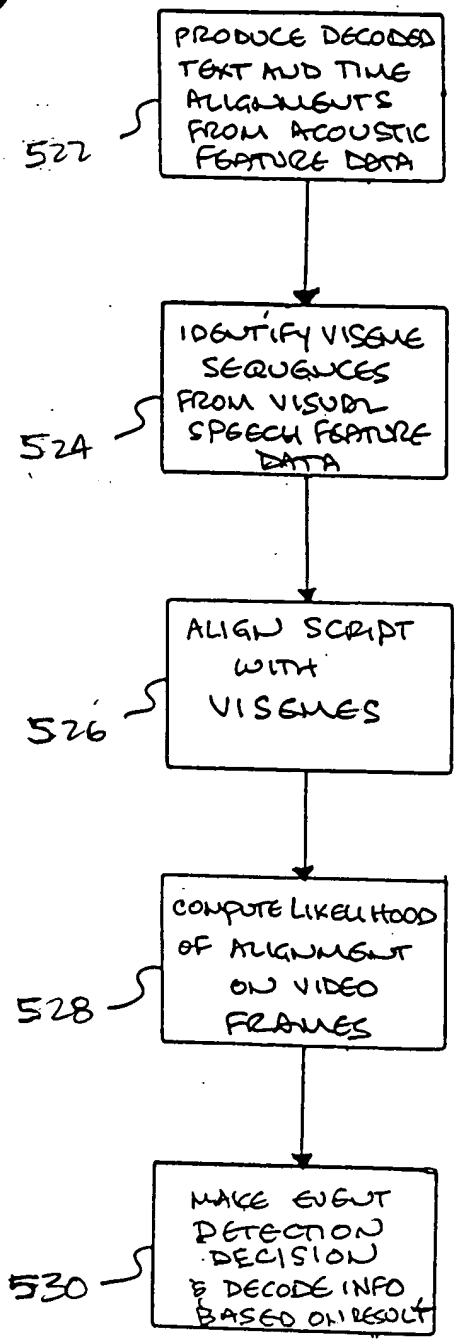
PRODUCE DECODED TEXT AND TIME ALIGNMENTS FROM ACOUSTIC FEATURE DATA
522
IDENTIFY VISEME SEQUENCES FROM VISUAL SPEECH FEATURE DATA
524
ALIGN SCRIPT WITH VISEMES
526
COMPUTE LIKELIHOOD OF ALIGNMENT ON VIDEO FRAMES
528
MAKE EVENT DETECTION DECISION & DECODE INFO BASED ON RESULT
530

FIG. 5D

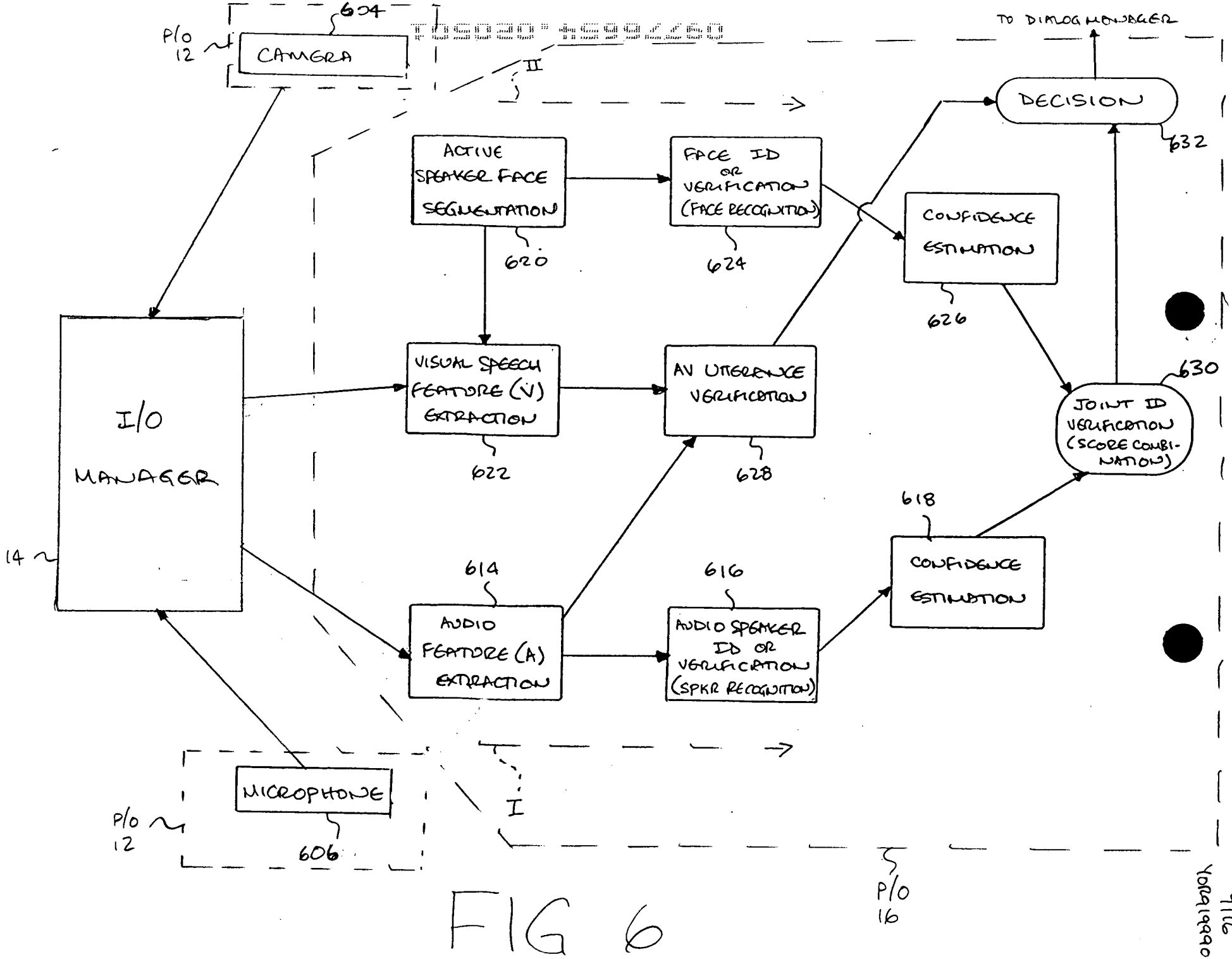
P/O 12
CAMERA
604
II
TO DIALOG MANAGER
DECISION
632
ACTIVE SPEAKER FACE SEGMENTATION
620
FACE ID OR VERIFICATION (FACE RECOGNITION)
624
CONFIDENCE ESTIMATION
626
I/O MANAGER
14
VISUAL SPEECH FEATURE (V) EXTRACTION
622
AV UTTERANCE VERIFICATION
628
630
JOINT ID VERIFICATION (SCORE COMBINATION)
618
614
616
AUDIO FEATURE (A) EXTRACTION
AUDIO SPEAKER ID OR VERIFICATION (SPKR RECOGNITION)
CONFIDENCE ESTIMATION
MICROPHONE
606
P/O 12
I
P/O 16

FIG 6

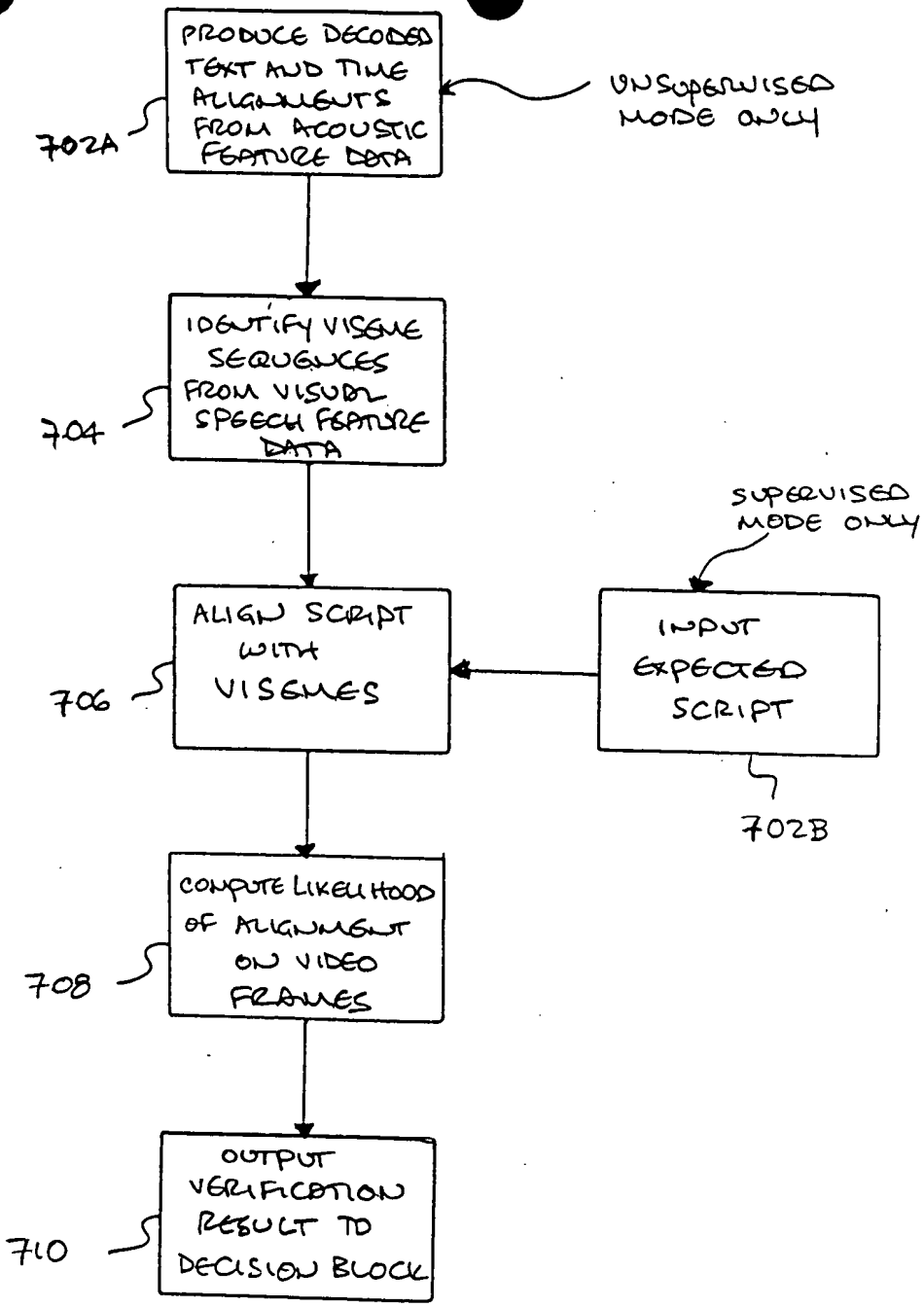
PRODUCE DECODED TEXT AND TIME ALIGNMENTS FROM ACOUSTIC FEATURE DATA
UNSUPERVISED MODE ONLY
702A
IDENTIFY VISEME SEQUENCES FROM VISUAL SPEECH FEATURE DATA
704
SUPERVISED MODE ONLY
ALIGN SCRIPT WITH VISEMES
706
INPUT EXPECTED SCRIPT
702B
COMPUTE LIKELIHOOD OF ALIGNMENT ON VIDEO FRAMES
708
OUTPUT VERIFICATION RESULT TO DECISION BLOCK
710

FIG. 7

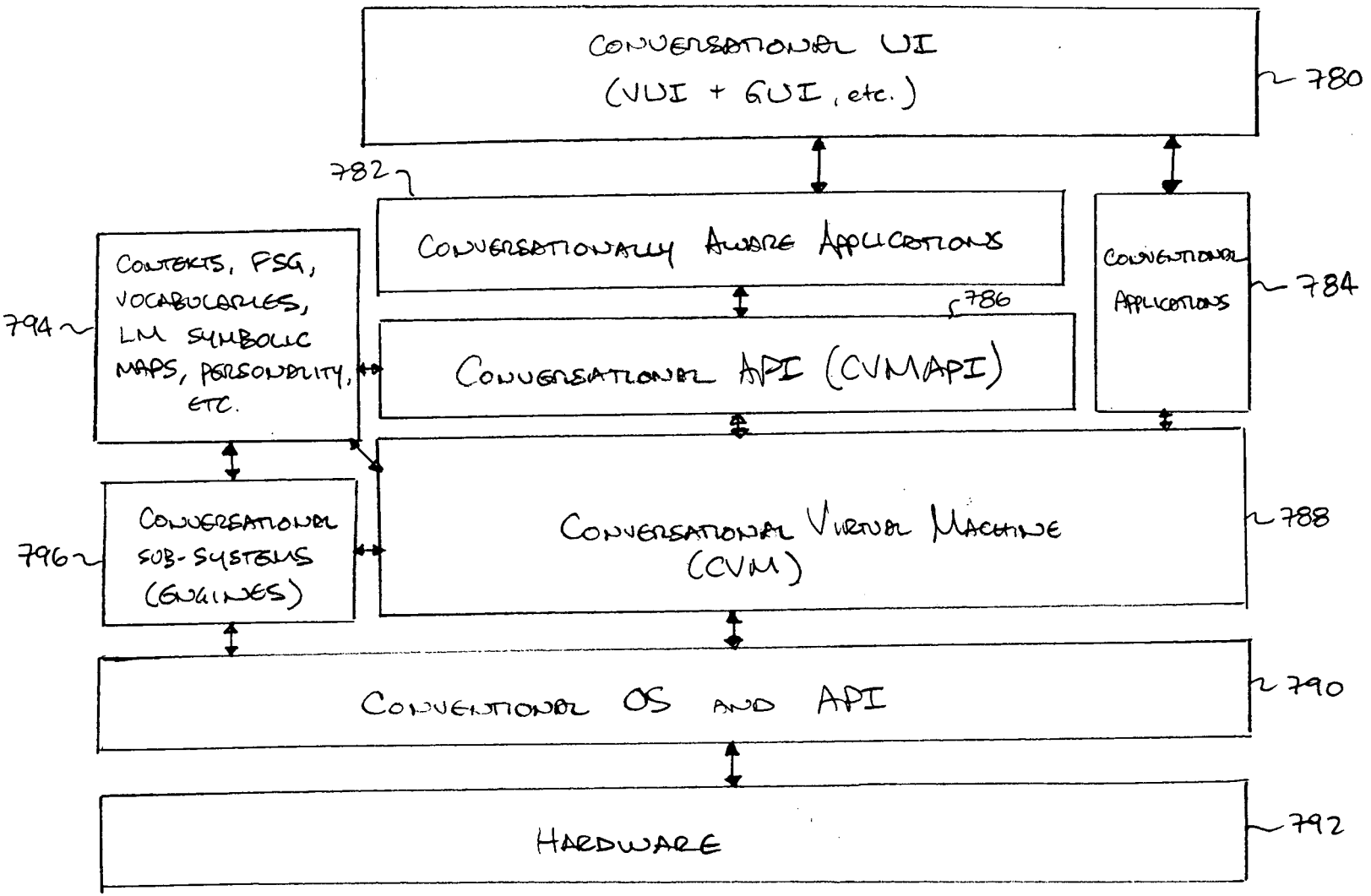
CONVERSATIONAL UI
(VUI + GUI, etc.)
780
782
CONVERSATIONALLY AWARE APPLICATIONS
786
CONVENTIONAL APPLICATIONS
784
CONTEXTS, FSG, VOCABULARIES, LM SYMBOLIC MAPS, PERSONALITY, ETC.
794
CONVERSATIONAL API (CVMAPI)
CONVERSATIONAL SUB-SYSTEMS (ENGINES)
796
CONVERSATIONAL VIRTUAL MACHINE (CVM)
788
CONVENTIONAL OS AND API
790
HARDWARE
792

FIG. 8A

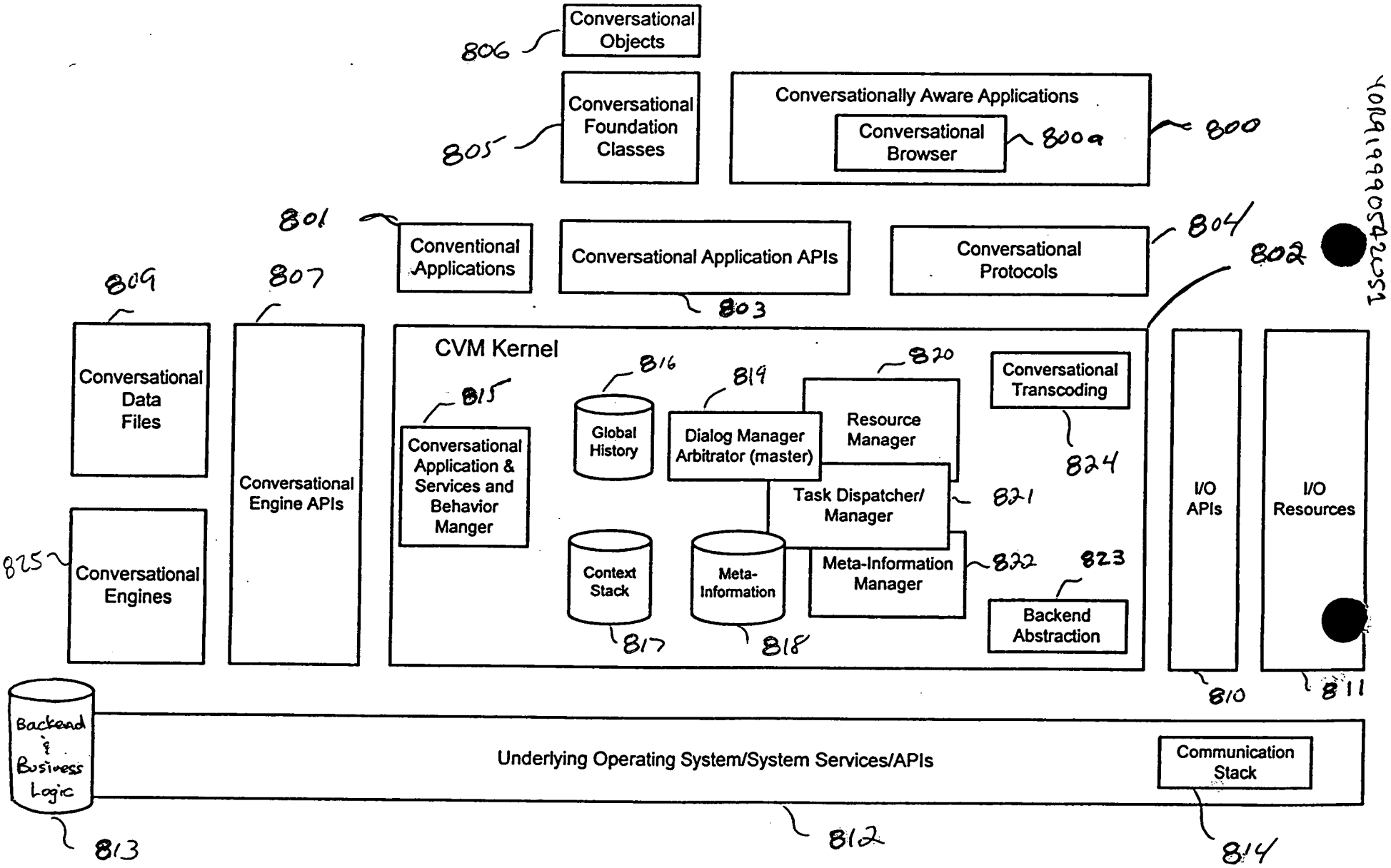
806
Conversational Objects
805
Conversational Foundation Classes
Conversationally Aware Applications
Conversational Browser
800a
800
801
Conventional Applications
Conversational Application APIs
803
Conversational Protocols
804
802
809
Conversational Data Files
807
Conversational Engine APIs
CVM Kernel
815
Conversational Application & Services and Behavior Manger
816
Global History
819
Dialog Manager Arbitrator (master)
820
Resource Manager
Conversational Transcoding
824
821
Task Dispatcher/ Manager
822
Meta-Information Manager
823
Backend Abstraction
Context Stack
817
Meta-Information
818
I/O APIs
810
I/O Resources
811
825
Conversational Engines
Backend & Business Logic
813
Underlying Operating System/System Services/APIs
812
Communication Stack
814

FIG. 8B

YOR919990542US1

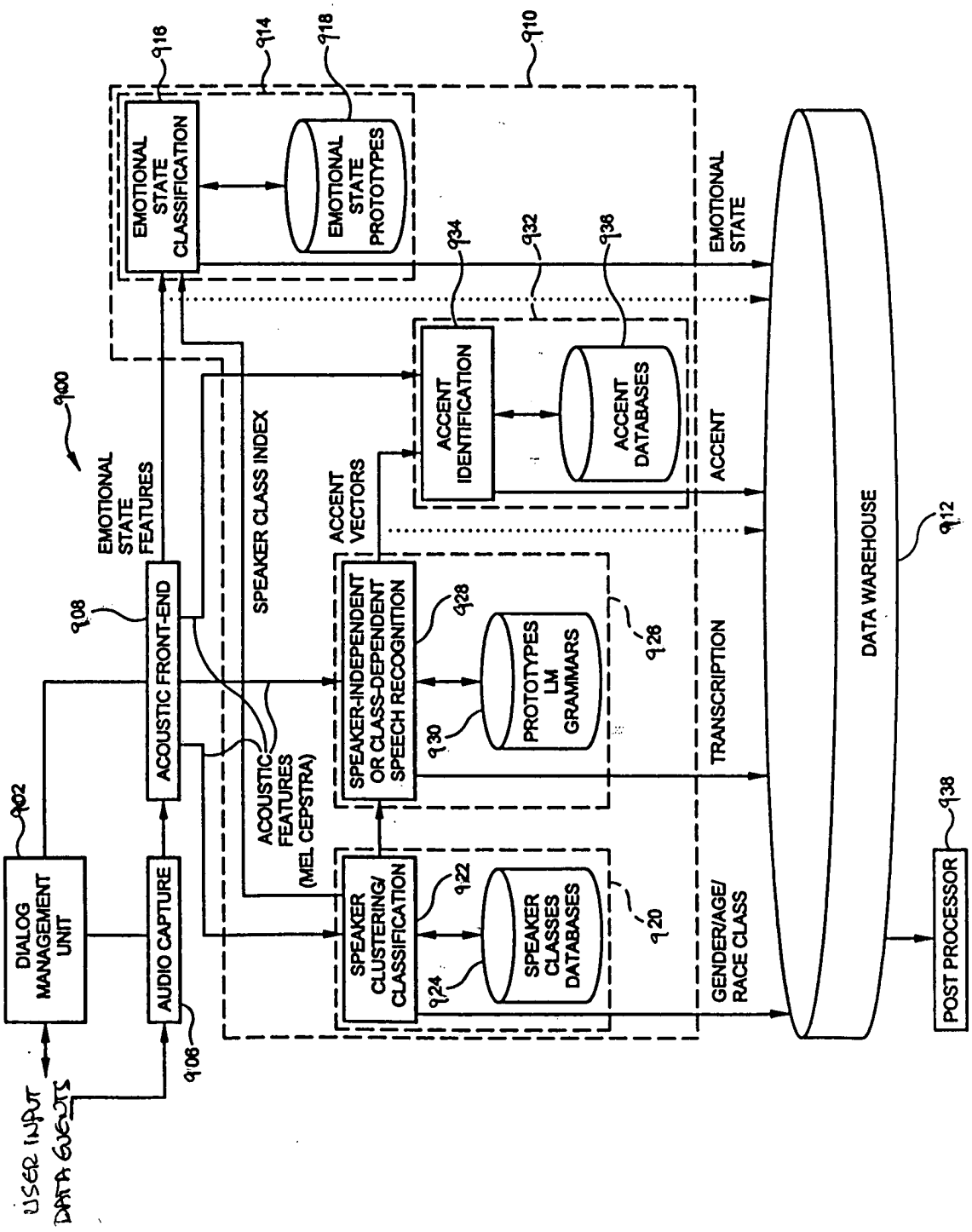

FIG. 9A

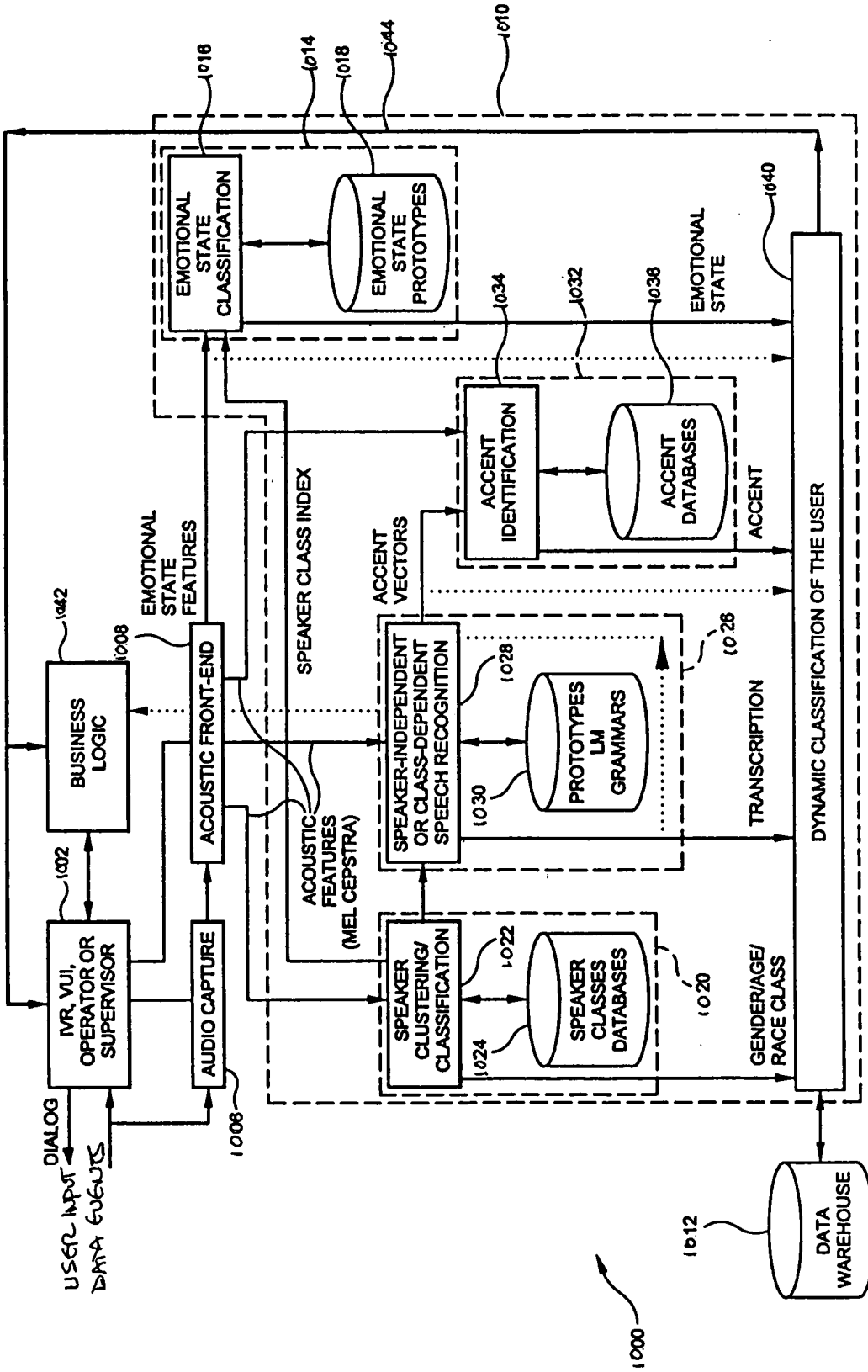
1000
1002
IVR, VUI, OPERATOR OR SUPERVISOR
USER INPUT
DIALOG
DATA OUTPUTS
AUDIO CAPTURE
1006
ACOUSTIC FRONT-END
1008
EMOTIONAL STATE FEATURES
ACOUSTIC FEATURES (MEL CEPSTRA)
SPEAKER CLASS INDEX
ACCENT VECTORS
BUSINESS LOGIC
1042
1010
1016
EMOTIONAL STATE CLASSIFICATION
1014
1018
1044
EMOTIONAL STATE PROTOTYPES
1034
1032
1038
EMOTIONAL STATE
1040
ACCENT IDENTIFICATION
ACCENT DATABASES
ACCENT
SPEAKER-INDEPENDENT OR CLASS-DEPENDENT SPEECH RECOGNITION
1028
1026
1030
PROTOTYPES LM GRAMMARS
TRANSCRIPTION
SPEAKER CLUSTERING/ CLASSIFICATION
1022
1024
SPEAKER CLASSES DATABASES
1020
GENDER/AGE/ RACE CLASS
DYNAMIC CLASSIFICATION OF THE USER
1012
DATA WAREHOUSE
POST PROCESSOR
1038

FIG. 9B

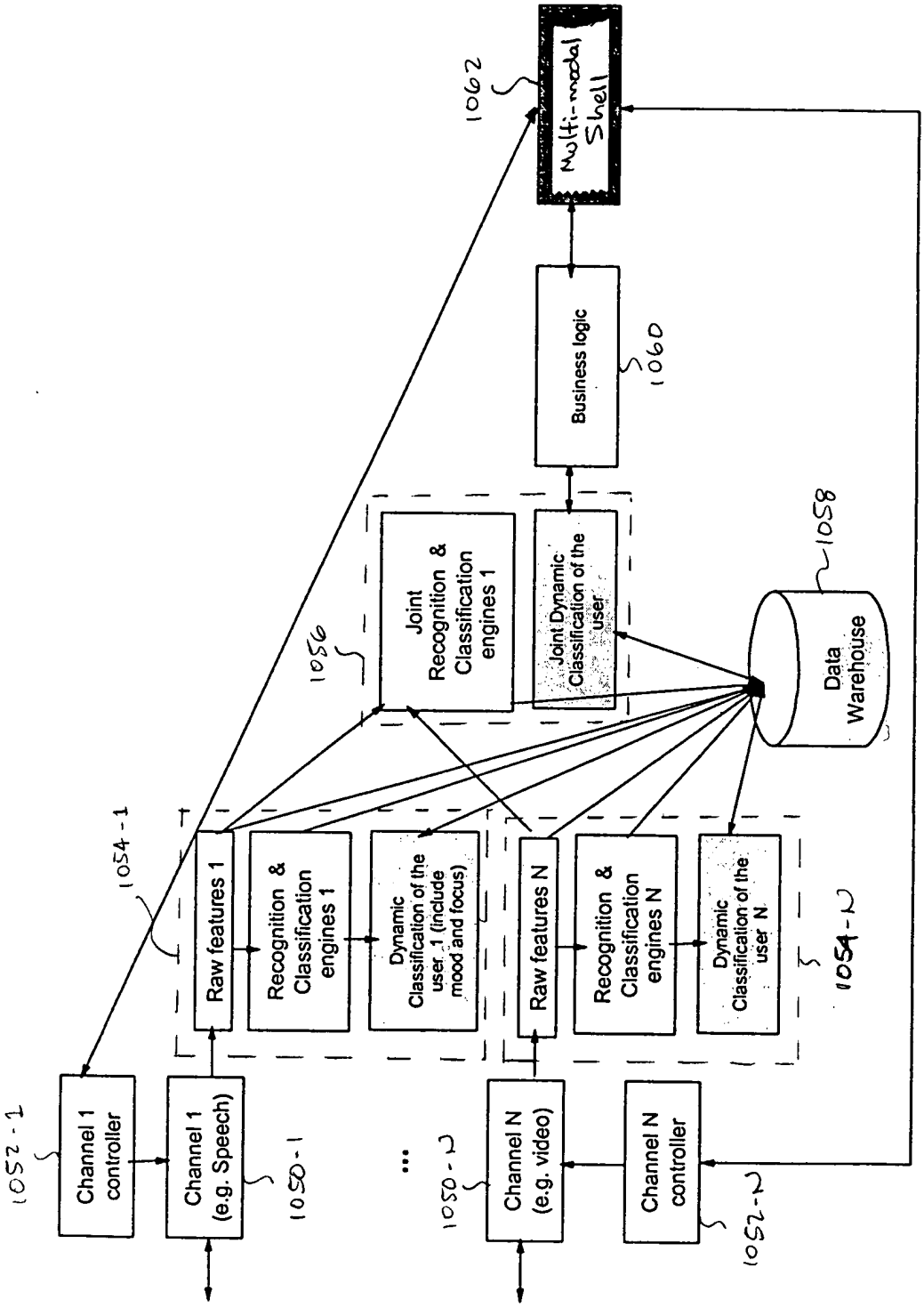
1062
Multi-modal Shell
Business logic
1060
1058
Data Warehouse
1056
Joint Recognition & Classification engines 1
Joint Dynamic Classification of the user
1054-1
Raw features 1
Recognition & Classification engines 1
Dynamic Classification of the user 1 (include mood and focus)
1054-N
Raw features N
Recognition & Classification engines N
Dynamic Classification of the user N
1052-1
Channel 1 controller
Channel 1 (e.g. Speech)
1050-1
1050-N
Channel N (e.g. video)
Channel N controller
1052-N

FIG. 9C

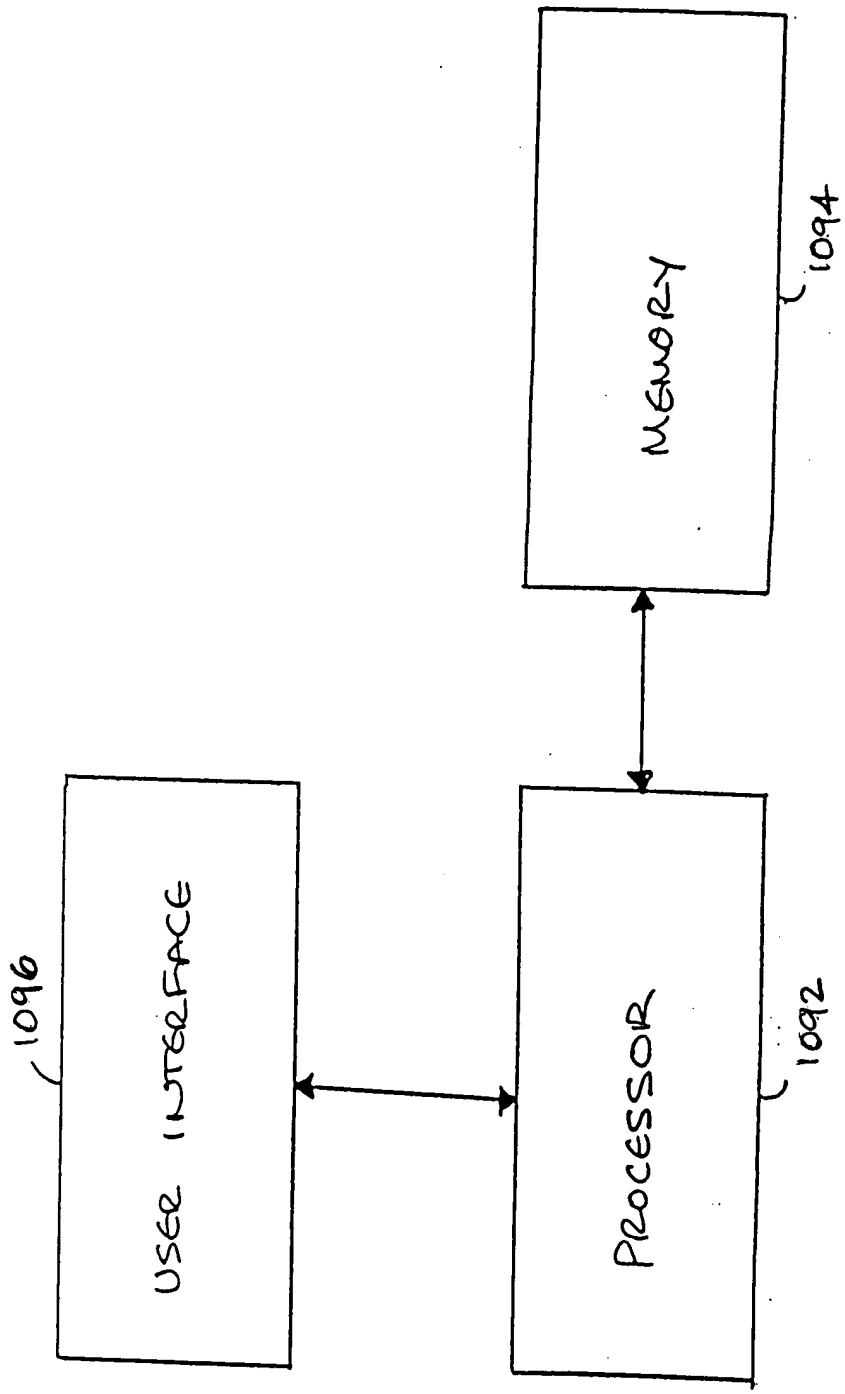

FIG. 10